ECA gesture strategies for robust SLDSs

Beatriz López[1] (student) and **Álvaro Hernández**[1] (student) and **David Pardo**[1] (student) and **Raúl Santos**[1] (student) and **María del Carmen Rodríguez**[2]

Abstract. This paper explores the use of embodied conversational agents (ECAs) to improve interaction with spoken language dialogue systems (SLDSs). For this purpose we have identified typical interaction problems with SLDSs and associated with each of them a particular ECA gesture or behaviour. User tests were carried out dividing the test users into two groups, each facing a different interaction metaphor (one with an ECA in the interface, and the other implemented only with voice). Our results suggest user frustration is lower when an ECA is present in the interface, and the dialogue flows more smoothly, partly due to the fact that users are better able to tell when they are expected to speak and whether the system has heard and understood. The users' overall perceptions regarding the system were also affected, and interaction seems to be more enjoyable with an ECA than without it.

1 INTRODUCTION

In this paper we examine certain complementarities of spoken dialogue and visual communication in Human-Machine Interaction (HMI). More specifically, we wish to identify effects of incorporating an animated agent onto a spoken language dialogue system (SLDS). Such dialoguing animated agents are commonly referred to in the literature as embodied conversational agents (ECAs) [1].

Our primary concern is to find benefits that may be gained by adding to a SLDS a visual channel of communication featuring an ECA. We are, of course, especially interested in improving aspects of human interaction with SLDSs that are particularly problematic. One major problem area is robustness. Dialogues often run into trouble for various reasons. For instance, when speech recognition errors occur it is usually difficult to recover from them. Error spirals are common [2], and even when the dialogue strategies designed specifically for error recovery are successful, interaction tends to become awkward, inefficient and "unnatural." Turn management is also tricky, and users are often not sure when they are supposed to speak.

What does an ECA bring into the picture? Most generally, a human-like figure adds a social element to the interaction. It may convey supra-linguistic information by performing gestures, including some designed as visual cues specifically to smoothen the flow of the dialogue making it seem more "natural" (for instance, by marking turn transitions), and others characterising expectations, mental processes (e.g., how well the system is understanding the user) and emotions (e.g., using emotional and empathic strategies to control user frustration when errors occur) - [3], [4], [5], [6], and [7].

[1] ETSIT Universidad Politecnica de Madrid, Spain; email: beatriz@gaps.ssr.upm.es
[2] Telefónica I+D, Spain; email: mcrg@tid.es

According to some critics, however, no real benefits of interaction with ECAs have ever been proved. ECAs, they add, can be misleading and create false expectations regarding the system's interactional and functional capabilities. Furthermore, they can be confusing, distracting, and even increase user anxiety and reduce the users' sense of control ([8], [9]).

As part of the research we are currently undertaking in the context of COMPANIONS -a European Union project [10]-, we have performed user tests on a dialogue system with and without an ECA in an attempt to isolate the effects of the ECA on the interaction. Our comparative analysis is focussed especially on finding gesture sequences that complement dialogue strategies designed to improve dialogue flow and robustness, thus resulting in improved overall interaction.

The application scenario we have designed is a domotic videotelephony service where users call "home" using mobile phones (simulated on a computer screen) to check the state of various home appliances. This task isn't important in itself in the scope of our experiment (here we are not especially interested in designing a real remote domotic control service); we use it solely to motivate dialogue that may go through the main stages identified in the literature for automatic dialogue generation [11].

This notwithstanding, remote domotic control applications are certainly interesting in their own right. Today new videotelephony applications are being developed for mobile terminals, gradually moving towards the use of directed spoken dialogue to access a variety of information services (like voicemail or videomail). Incorporating ECAs onto this new visual channel affords challenges of its own. For instance, screen space is more limited, so what ECA size, appearance and gestures are best and whether it is appropriate to have an ECA on screen in the first place are all relevant questions for research.

The rest of the article is organized as follows: Section 2 presents the dialogue strategies we have implemented to increase robustness and the ECA behavioural schemes we have associated with them. Section 3 describes the types of ECA parameters considered in our evaluation. In Section 4 we explain how the empirical test was set up, and we show its structure. Section 5 shows the main results of the experiment, with discussion. Section 6 brings together the main findings and anticipates the next steps of research.

2 DIALOGUE AND GESTURE STRATEGIES

Among the more critical dialogue situations for which it is worth examining the positive effects an ECA could have are the following:

- Turn management: Here the body language and expressiveness of agents could be exploited to help regulate the flow of the dialogue [12]. Usability experimental analysis on how the facial feedback provided by avatars can

make turn-taking smoother in the COMIC multimodal dialogue system has been presented in [13].

- Error recovery: The process of recognition-error recovery typically leads to a certain degree of user frustration (see [14]). In fact, once an error occurs it is common to enter an error spiral, because as the user becomes increasingly frustrated, her frustration leads to more recognition errors, making the situation worse [15]. ECAs may help to limit such feelings of frustration and by so doing make error recovery more effective [16].
- User confusion: A common problem in dialogue systems is that the user isn't sure what the system is doing and whether or not the dialogue process is working normally [17]. This sometimes leads the dialogue to error states that could be avoided. The expressive capacity of ECAs could be used to help the user keep track of what stage the dialogue is in (i.e., what the system is doing and expecting from the user).

We have designed a dialogue strategy to deal with various critical dialogue stages, react to different recognition confidence levels and manage error situations. Associated with the dialogue strategy is an ECA gesture scheme, with a set of gestures corresponding to each dialogue stage. Table 1 shows each dialogue stage, what prompts it, and the associated ECA behaviour. The gesture repertoire of our ECA is partially based on relevant gestures described in [1] and [12], and on recommendations in [18], [19], [20], [21], and [22], to which we have added a few suggestions of our own.

Aiming to define ECA behaviour during the interaction, we have tried to exploit the following supra-linguistic resources: conversational skills (such as beat gestures to emphasize information, nodding and "don't understand" gestures, waiting posture, etc.), shifts in camera shots and lighting intensity (in order to create "proxemic" effects that might be meaningful to the user), and the recreation of an empathic attitude in the ECA (smiling or offering an expression of apology) to try to keep user frustration low when interaction problems occur.

In the rest of this section we explain in a little more detail the dialogue-gesture scheme for each stage summarised in Table 1.

Initiation. Upon first encountering an ECA the user may "humanise" the system [23] and expect from it a lot more than it is actually capable of. Users may tend to speak with less restraint, making it more difficult for the system to understand them. The end result is likely to be somewhat disappointing and frustrating. Another possible effect we should consider is that contact with a dialoguing animated character may have the effect that the user's level of attention to the actual information that is being given is reduced ([24], [25]), especially in the case of new users (as our test users are). Thus, the goal at initiation is to present a human-like interface that is upon first contact less striking and less distracting, and one that clearly "lays down rules" of the interaction and sets the user on a track that is tightly focussed on the task at hand.

In order to try to foster a sense of ease in the user and help her focus we have designed a welcome gesture for our ECA based on the recommendations in Kendon [20], (see Table 1).

Termination. It is confusing if a dialogue concludes without the user being aware of it. It is important to end with a clear farewell message. We have complemented this with typical farewell gestures in human-human interaction [1].

Table 1. Gesture repertoire for the main dialogue stages

MAIN DIALOGUE		
Dialogue stage	**Description (when it occurs)**	**ECA behaviour (movements, gestures and other cues)**
Initiation	At the beginning of the dialogue	Look straight at the camera, smile, wave hand. Zoom in for task explanation. Zoom out, lights dim.
Turn management	*Take Turn:* when the system starts to speak	Look straight at the camera, raise hand into gesture space. Camera zooms in. Light gets brighter.
	Give Turn: when the system prepares to listen to the user	Look straight at the camera, raise eyebrows. Camera zooms out. Lights dim.
Wait	When a timeout occurs	Slight leaning back, one arm crossed and the other touching the cheek. Shift of body weight.
Help	When the system gives some explanation to the user	Beat gesture with the hands. Change of posture.
Confirma-tion (low confidence)	When the system cannot understand something the user has said.	Slight leaning of the head to one side, stop smiling, mildly squint.
Confirma-tion (high confidence)	The system has recognised the user utterance with a high level of certainty	Nod gesture, smile, eyes fully open.
Acknowle-dgement of misunders-tanding	After user informs the system that it has misunderstood what he or she has said. Speech: a) apology; b) repetition or rephrase request	Apology: Head aside, raise inner eyebrow central, head down, eyebrow of sadness (to show remorse). Request: Show expression of interest by opening eyes, and smiling slightly.
Error recovery with correction	When the user has corrected a recognition error and the system confirms the correction	Lean towards the camera, beat gesture.
Termination	Goal: to show that the dialogue is being closed. Speech: farewell message.	Looks straight at the camera, nod, smile, wave hand.

Turn management. Turn management involves two basic actions: taking turn and giving turn. Dialogue fluency improves and fewer errors occur if alternate system and user turns flow in orderly succession with the user knowing when it is her turn to speak. It is important to point out that we have not allowed barge-in (i.e., the user cannot interrupt the system because the system doesn't listen to the user –the speech recogniser is inactive– while the system is speaking). This makes for a less flexible dialogue scheme than may be generally desirable, but we hope it offers at least two advantages: firstly, in certain problem situations such as error spirals [26] it may well be most advisable never to allow the user to interrupt while the system is trying to reach a stable, mutually understood dialogue state. Since these are the cases we are most interested in, it makes sense to work with barge-in-free dialogue. Secondly, if users try to speak when they're not "supposed to" (our users are not told

they cannot interrupt the system) this usually leads to no-inputs (when the system isn't aware that the user has said something), no-matches (the system is unable to understand the incomplete utterance it "hears"), and perhaps recognition errors. Turn management then becomes more critical, and the consequences of confusion regarding who's turn it is more obvious. Thus the role an ECA may play in clarifing turn possession and turn transitions should be more apparent.

Our ECA strategy is as follows: When it's the ECA's turn to speak the camera zooms-in slightly and the light becomes brighter; while the ECA is approaching it raises a hand into the gesture space to "announce" that it is going to speak (see Figure 1). When it's the user's turn the camera zooms out, the lights dim and then the ECA raises its eyebrows to invite the user to speak. The idea is that, hopefully, the user will associate different gestures, camera shots and levels of light intensity with each of the turn modes.

Figure 1. Visual sequence of turn transition from user to ECA.

Confirmation. Once the user utterance has been recognised, information confirmation strategies are commonly used in dialogue systems. Different strategies are taken depending on the level of confidence in the correctness of the user locution as captured by the speech recognition unit [22]. Our dialogue scheme and the associated gestural strategies are as follows:

- *High confidence in recognition:* The dialogue continues without confirmation request. The ECA nods her head [1], smiles and opens her eyes wide to show the user that everything is going well and the system understands her.
- *Intermediate confidence:* The result is regarded as uncertain and the system tries implicit confirmation (by including the uncertain piece of information in a question about something else). This allows the user to correct the system if an error did occur, and to feel everything is going well if what the system understood was correct. No specific ECA gesture was designed for this case. The idea is to keep the user speaking normally and without hyperarticulating (which would make recognition more difficult [15]).
- *Low confidence:* The dialogue becomes more guided with the system asking the user to repeat or rephrase. The ECA leans her head slightly to one side, stops smiling and mildly squints (a "what was that you said?" gesture; see Figure 2).

Acknowledgement of misunderstanding. A particularly delicate situation arises when the system misunderstands the user. If the user tries to correct the system or point out that it has misunderstood, the system will hopefully realise what has happened. It then tries to keep the user in a positive attitude and avoid her distrust while seeking to obtain the correct information. The dialogue scheme to pursue this consists in an apology followed by a kind request for a repetition or rephrase. The ECA gestures accordingly (see Table 1), stressing the system's "interest" in getting it right to further motivate the user and preserve her trust.

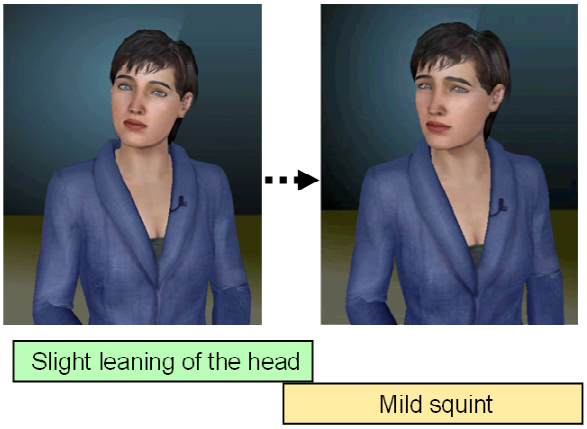

Figure 2. ECA gesture sequence expressing low confidence in the comprehension of the user's utterance.

Error recovery with correction. If the user says that recognition errors have taken place and gives the correct information at the same time, the ECA repeats the corrected information by leaning towards the camera and marking the words by means of beat gestures with both hands.

Help. A help message is given either when the user requests it or when the system has failed to hear the user say anything for longer than a reasonable waiting period. The ECA emphasizes the more important information in the help message with beat gestures performed with the hands. The idea is to see whether this captures the interest of the user, makes her more confident and the experience more pleasant, or if, on the contrary, it is distracting and makes help delivery less effective.

Wait. As we discussed before, it sometimes happens that the user doesn't realize it is her turn to speak. To help the user realise the system is waiting for her to say something the ECA performs a waiting gesture: leaning back slightly with her arms crossed and shifting the body weight from one leg to another.

3 EVALUATION PARAMETERS

As was mentioned in the introduction, our main goal is to evaluate how well ECAs can work in improving HMI performance parameters and user satisfaction. The approach we have taken is based on Möller et al.'s taxonomy of quality factors for dialogue systems [27] and the ITU P.851 recommendation [28] on evaluating dialogue systems, to which we have added questions as we have seen appropriate to evaluate user perceptions related to the ECA. We combine the system and interaction, performance and event data registered automatically, with user's responses to questionnaires (note that although recognition performance data is interesting, the goal of the experiment reported in this paper is not to evaluate how well the speech recogniser works).

In order to measure the influence of an ECA on user satisfaction we have compared a dialogue system that includes an ECA in the interface with one without ECA along a range of user-centred parameters. These parameters fall into three classes:

- *Typical dialogue system parameters* (automatically collected in the questionnaires) covering aspects of system performance, dialogue flow, information offered by the system, usefulness and overall evaluation, overall impression and perception of task success.
- *Impressions felt while using the system*: User emotions ("relaxed", "confident", "happy", "bored", "dejected", "angry" and "clumsy") and sensations ("pleasant", "fun", "interesting", "frustrating", "confusing" and whether they were surprised by anything). Participants were also invited to write comments about different aspects of the system.
- *Specific questions concerning the ECA* regarding both gesture design (clarity, naturalness, range of the gesture repertoire) and the perceived personality of the ECA ("expressive", "likable", "polite", and how comfortable it is to speak with the agent).

We have also added a time dimension to see whether we can determine how users' expectations evolve through use of the system (other studies, such as that in [29], have focused primarily on user expectations). We do this by repeating certain questions at different stages of the test.

4 EXPERIMENTAL SETUP

4.1 System implementation

The architecture of the test environment is based on web technology, with which we simulate a mobile phone interface. Figure 3 shows the two different interaction scenarios we have compared: one (on the right) corresponding to what we have called the ECA metaphor scenario, and the other (on the left) with a still image (representing "home") that we call VOICE metaphor scenario (SLDS without ECA). Different users interact with these two different scenarios providing contrastive experimental data that will allow us to evaluate the ECA metaphor vs. the VOICE metaphor. The system is implemented on a web page that contains two frames. In the left frame there is a column of labels that show the test user what stage of testing he or she is (not to be confused with the dialogue stage which is not indicated). The main interface is displayed in the right frame and shows a mobile phone running a videotelephony application. Tactile interaction is not active at any stage.

Figure 3. Interface displays for VOICE (left) and ECA (right) metaphors.

All the contents of the evaluation are hosted on an Apache Tomcat web server. Throughout the test, users face a series of evaluation questionnaires and experimental dialogue interactions. The questionnaires are implemented using HTML forms, and the information collected on them is transferred to JSP files and then stored in a database. Our test environment uses Nuance Communications' speech recognition technology [30]. The ECA character was created by Haptek [31]. The dialogues are implemented with Java Applet technology, and they are all packed and signed to guarantee fast download and access to the audio resources. Dialogue dynamics are programmed. Nuance's speech recognition engine provides a useful Java API that allows access to different grammars and adjusting a range of parameters depending on the characteristics of each application. Finally, interaction parameters (such as utterance durations, number of turns, number of recognition errors, etc.) are recorded automatically during the test interactions.

4.2 Description of the experiment

Testing was done in a small meeting room. Each user was sat at the head of a long table in front of a 15" screen. Two different views of the user interacting with the system where video-recorded to provide us with visual data to inspect and annotate the subject's behaviour. A frontal view was taken from the top edge of the user's screen, and a lateral view was recorded from a wide-angle position to the right of the user. Both views were taken with Logitech Quickcam Pro 4000 webcams. The users interacted with the system using a headset microphone, and the system prompts are played through two small speakers. Half of the users interact with a system only through spoken dialogue; the other half encountered an interface that includes an ECA. All user-system dialogue was in Spanish.

The evaluation was designed so that users could carry out the test with minimal intervention on the part of experimenter. The stages of the test were as follows:

1) Brief explanation: An experimenter briefly explains to each test participant what the general purpose (to "evaluate automatic dialogue systems") and methodology of the test are, as well as the tasks that lie ahead. We try to emphasize the importance of the answers given in the questionnaires.

2) Opening questionnaire to learn about the users' prior experience and expectations.

3) Training and verification phase (and associated questionnaires): Users are asked to enrol in a secure access application using voice recognition, and then to verify their identity. The interaction method is a rigid, directed dialogue, with an ECA for half of the users. We will not deal with this aspect of our test in this paper. We mention it here for two reasons: the first is for completeness and accuracy of our account of the testing procedure; the second is that some questions contained in the questionnaires at this stage are repeated later in the final questionnaire, which enables us to analyse how users' opinions and expectations evolve throughout the test.

4) Dialogue phase: Users are given three dialogue tasks. In each task users are asked to find out the state (on/off) of a household device ("the bathroom lights", "the fan in the bedroom", and "the living-room television set". The automatic speech recogniser and the dialogue system function freely (i.e., they are not programmed to give certain answers; it is a real working

system). Half of the users interact with an ECA and the other half without.

5) Final questionnaire to get the user's overall impression of the system, its main elements and the more important aspects of using it. As mentioned before, some questions are the same as in previous questionnaires to provide information regarding the evolution of users' perceptions throughout the various stages of system use.

5 EXPERIMENTAL RESULTS

We carried out testing with 16 undergraduate and graduate students (7 female and 9 male), of ages ranging from 19 to 33, divided into two groups (8 users in each group), one to test the system with the ECA interaction metaphor and the other with the VOICE metaphor.

Our analysis is mainly based on the users' answers to the questionnaires and on the performance parameters registered during the course of the user-system interaction. As previously mentioned, we based our questionnaire design partly on the ITU-T P.851 recommendation [28], which identifies a variety of conceptual dimensions or categories that should be taken into account when writing questions to evaluate users' opinions on a comprehensive range of aspects related to quality of interaction with dialogue systems. We have added similar question categories that deal with ECA presence and gestures, and also a set of questions to inquire about the users' emotions while using the system.

We have reached the following results by a) comparing the performance and the answers to the questionnaires of the ECA metaphor group of users with those of the VOICE metaphor group; and b) analysing how performance and responses to certain questions evolve throughout the test. In addition, we have looked at users' comments, given at certain points in the questionnaires, and compared them to the findings in a) and b).

We carried out a series of two sample t-tests, setting the significance level at 5% (p=0.05). Questionnaire responses were collected on Likert-type 5-point response formats.

In the rest of this section we present the main findings obtained from these comparative analyses.

5.1 Sensitivity to errors and user frustration

We found some statistically significant differences between ECA and VOICE metaphors regarding certain factors related with robustness in difficult dialog situations (e.g., when the system acknowledges having misunderstood something, or when the system doesn't "get" what the user has said). Specifically:

Average user awareness of system recognition errors is lower for ECA users. In spite of the fact that the minor difference we found in the actual average numbers of recognition errors between both of the tested interaction metaphors was not statistically significant, there was a striking, statistically significant, difference in the answers to the question *"Did the system make many mistakes?"* (1- very many ... 5 - none): a mean value of 3.8 for the ECA metaphor vs. 2.6 for the VOICE metaphor ($t(12)=3.16$; p-value=0.004). User frustration while interacting with the system was also markedly lower for the ECA group, as indicated by the 1.4 (ECA) vs. 2.6 (VOICE) mean values ($t(9)=-2.52$; p-value=0.016) of the responses to the question: *"Was the experience [of using the system] frustrating?"* (1 – no, not at all ... 5 – yes, very much so).

The measured differences in the two previous parameters between the ECA and VOICE scenarios possibly reflect relevant advantages, at least in terms of how it affects user perception, of the use of ECAs with appropriately designed gestures, both to deal with problematic dialog stages such as error recovery situations and to provide users with visual cues of how well the system is understanding her (i.e., with what level of confidence; see Section 2). We could be seeing here a variant of the persona effect [32], a phenomenon widely reported in the literature according to which users tend to perceive a particular task as easier when they interact with an ECA in order to carry it out, without there being any real improvement in performance (success in task execution and efficiency) when compared with users doing the same without an ECA. In our case no significant difference was found between the two test groups regarding perception of ease of use. However, believing the system made fewer mistakes could be a related effect.

There may be more to it, though, and user frustration and perception of performance quality may be linked to actual improvements in dialogue flow and in the users' knowledge of what is going on (what the system is doing and expecting the user to do). We now turn to exploring these possibilities briefly.

5.2 Dialogue coordination and fluency with visual cues for turn-switching

Efficiency and fluency of interaction are important factors (identified in [28]) in which we have also found differences between the ECA and VOICE metaphors. Users' perception that *"Dialoguing with the system led quickly to solve the task proposed"* (1 - totally disagree ... 5 - totally agree) was on average greater in the ECA group (4.2) than in the VOICE group (3.2) ($t(12)=3,16$; p-value=0.004).

This is not simply, or not solely, a subjective impression induced by the presence of the ECA, which would make it an instance of the persona effect. In fact, a close examination of our experimental ECA-supported dialogues shows that users easily learn when they are supposed to speak to the system (i.e., when it is their turn). This helps prevent most of the typically observed failed barge-in attempts and time-outs, which we found occurred more often for our VOICE metaphor users. Some of these users said they had felt confused at certain stages of the dialogue (e.g., *"between tasks there were silences and I didn't know if I was supposed to say anything," "a couple of times I think I spoke too early and that's why the system didn't get what I said," "it would be better if some sort of visual sign told you when the system is ready to listen"*).

However, we found no statistically significant differences between the two groups of users as regards task duration and number of turns taken, which are, of course, two important efficiency indicators. This notwithstanding, all of the main performance indicators were slightly better for the ECA group than for the VOICE group: average dialogue duration (μ_{ECA}=38655ms (std=18688); μ_{Voice}=47657ms (std=34043)), total duration of user turns (μ_{ECA}=4267ms (std=1745); μ_{Voice}=6182ms (std=4615)), number of dialogue turns (μ_{ECA}=5.70 (std=2.17); μ_{Voice}=6.33 (std=4.43)), number of time-outs turns (μ_{ECA}=0.08 (std=0.40); μ_{Voice}=0.20 (std=0.72)), number of times a help message is given (when the system

"realises" the user may be in trouble or confused) (μ_{ECA}=0.04 (std=0.20); μ_{Voice}=0.12 (std=0.44)) and number of no-matches (μ_{ECA}=0.25 (std=0.53); μ_{Voice}=0.41 (std=0.50)) were all lower for the ECA group.

Our sample sizes are rather small so we need to increase them to see if these observed differences become statistically significant. But for the time being it is reasonable to interpret our findings as possible evidence of a combination of a persona effect with the fact that ECA-metaphor users learn how to interact with the system more easily and feel more in control, and actually achieve a more coordinated dialogue (if not significantly more efficient in terms of time) than VOICE-metaphor users.

Thus, it seems our visual feedback channel featuring an ECA displaying contextual dialogue management cues may be providing supra-linguistic information that users are able to interpret correctly, which translates into an improved coordination, which in turn increases the users' impression of the dialogue being fast, efficient and under control. This could also be related with our finding that user perception of system mistakes and user frustration were lower for the ECA group, as reported above.

But what are these visual cues that appear to be so useful? Our findings allow us to suggest that the visual information strategy for turn-switching that we have implemented –involving a combination of gestures and lighting and camera zoom effects– may be creating a "proxemic-code" that helps avoid the complicated, problem-laden interaction patterns reported in [13], where user-ECA interaction suffers from rather severe coordination problems. Furthermore, our proxemic strategy is as simple as the good old invitation to speak using beeps, with the advantage that in our tests we haven't observed any sign of rejection as may arise with the use of artificial-sounding beeps. Users seem to accept meaningful proxemic shifts as a "natural" part of dialogue interaction.

5.3 User expectations and perception of dialogue capability

The users' impression of how powerful the system's dialogue capabilities are, combined with the users' expectations regarding these capabilities, has an important impact on the users' overall assessment of the system [28]. Our experimental results show that the ECA-metaphor group was impressed with the system dialogue capability, although somewhat less than the VOICE-metaphor group, the former grading with an average of 3.9, and the latter 4.5 (3.0 being the neutral score), on the question: *"Were you positively or negatively surprised by the system's dialogue capability?"* (1 - very negatively surprised … 5 - very positively surprised) (t(13)=-2.12; p-value=0.027). This result is in agreement with the findings in other research efforts (see, e.g., [10]).

A plausible explanation has to do with the effect, discussed in Section 2, by which users that encounter an "embodied" interface tend to be overoptimistic with regard to the system's capabilities, assuming these to be more on a par with those of human beings. But, since in fact we have the same dialogue engine behind both our ECA and VOICE-metaphor interfaces, users of the former tend to end up being less impressed with the system's conversational skills –having expected more but getting the same– than users of the latter.

This, of course, notwithstanding the fact that the users in the ECA group don't really "get the same," if we consider that, on average, they experience a smoother dialogue, as we saw previously. The following qualitative impressions expressed by our test users may add a little perspective to the analysis:

"In the beginning my main feeling was one of mistrust because it was a new experience, but afterwards it was pleasant and it was very easy to become accustomed to it."

"I thought that the interaction with the system would be less comfortable, but the system understood me very well."

Here we see that initial expectations might not be so positive after all, and that the experience of interacting with the system did in fact exceed at least some of the users' expectations. We clearly need to carry out further tests to shed light on the intricacies of user expectations and their evolution through system use.

5.4 Emotions

Apart from frustration, the only other feeling for which our data shows a statistically significant difference between the ECA and the VOICE group is happiness (users in both groups felt similarly relaxed, confident, bored, dejected, angry and clumsy, for instance). The ECA group averaged 4.0, against 3.1 for the VOICE group, in their replies to the question: *"While you were interacting with the system, did you feel happy?"* (1 - no, not at all … 5 - yes, very much so) (t(13)=1.99; p-value=0.034).

It is clear that the observed difference in emotional response between the two test groups, favouring as it may the use of an ECA, was only very slight. After all, the whole experimental procedure is short and fairly simple, and test users have very little at stake performing the test, so it seems unlikely that strong emotional responses might appear. However, in future experiments we plan to design longer, more complex tasks and, by increasing the sample size, we hope to be able to determine more precisely how our ECA affects user emotions, if at all, and how these might affect overall usability and user acceptance.

5.5 ECA expressiveness

We invited the test users to give us their views regarding the ECA's gestures and expressiveness. These are a few revealing samples:

"I very much liked the expressiveness of the animations."

"I found the agent and the agent's gestures surprising."

"The face gestures were very well designed, but the hand gestures could distract you."

"I liked the ECAs very much. They're very funny."

These opinions are encouraging, especially as there are studies that point out that in order to improve the believability and naturalness of an ECA it is essential to give it a consistent personality and to make it expressive (see, e.g., [33]).

Furthermore, in our study we have observed that the users' opinion of the ECA's expressiveness increases with use after first contact (which occurs in the identity verification phase of the test): the average score for *"Is the agent expressive?"* (1 - no, absolutely not … 5 - yes, very much so) increased from 3.5 after first contact to 4.1 at the end of the test (t-value=-3.42; p-value=0.006). Similarly, users' impression of ECA friendliness (another relevant factor connected to user expectations; see [34])

also increases slightly with use, from 4.1 to 4.5 (t-value=-2.05; p-value=0.040).

Expressiveness and friendliness may be "humanising" the ECA [35], but in a way that, rather than leading ultimately to disappointment, keeps users in a positive attitude and raises their interest in a natural-feeling interaction. This happens though the course of time (the little time our test lasts), which may be yet another piece of evidence that our ECA doesn't trigger unrealistic expectations upon first appearance, but gradually "wins users over."

Finally, we mention that in the present work we have not focused on specific gesture design (which gestures were preferred, which were perceived as being the clearest, and so on). However, prior to the present experiment we carried out a successful gesture validation test on the repertoire displayed by our ECA [36]. The comparative experiment discussed in this paper also serves as *implicit* overall gestural validation thanks to the interaction improvements we have observed. By analysing the video recordings of the user tests (which we will do shortly) we hope to obtain deeper insights on the effects of specific gestures –especially those we have designed with a view to improving dialogue robustness in difficult situations– and on how we might refine them.

6 CONCLUSIONS AND FUTURE WORK

Our line of research is intended to help make some progress in identifying the pros and cons of Embodied Conversational Agents (ECAs). In this article we have presented a research scheme in which we consider the main problem situations that typically arise in automatic dialogue generation. In order to improve the robustness and the ease-of-flow of the dialogue we have implemented a gesture repertoire to be displayed by an ECA at each stage of the dialogue. These gestures are designed to convey to the users meaningful supra-linguistic information regarding the state of the dialogue throughout the interaction, and to try to keep the user in a positive frame of mind. We have proposed evaluating how well these strategies work by setting up an experiment to compare two interaction scenarios or metaphors (ECA metaphor vs. VOICE metaphor).

We found that the ECA contributed to keeping user frustration low, especially when recognition errors occurred (which is the most delicate scenario). This result suggests that the error management strategies employed are working, particularly: a) implicit confirmation with no ECA reaction when confidence in recognition is intermediate; b) performing a "What was that you said?"-type gesture to show the user the system isn't sure it has understood but is making an effort to (when confidence in recognition is low); and c) acknowledging misunderstandings with an apology and an accompanying gesture sequence to reassure the user that the system knows what has happened and is trying to put things right.

Also worth mentioning is the observed improvement in dialogue fluency (especially in connection with turn changes) with the ECA interaction metaphor. The combined use of specific gestures and proxemic effects (playing with "camera" shot distance and light intensity) seems to be a promising alternative to the traditional 'beep' signal. In the absence of acoustic signals or visual cues, some users start speaking before the system is ready to listen. When visual cues are added, however, users display a greater tendency to wait until they see the animated figure is inviting them to speak. These strategies add naturalness and smoothness to the flow of the interaction.

On the negative side, the ECA's human-like appearance could potentially cause users to ultimately be somewhat disappointed with the system's dialogue ability, probably because of the false expectations such an appearance gives rise to, as has already been reported in the literature. Our results cannot confirm nor disprove this effect. However, we have seen indications that our ECA doesn't generate expectations in users that are too far off the mark. Indeed, users seem to appreciate the ECA more after they have interacted with it for a while. Nevertheless, this is an area we must examine more closely in future work.

The signs on which we have based our observations are only mild. We will continue testing with this experimental set-up, after which we will analyse all the gathered information, including the video recordings, to confirm (we hope) the effects reported in this paper and to refine our findings and discover more relationships between the interaction aspects we have considered (what we have presented here is a first batch of results that don't fully exploit the possibilities of the dialogue and gesture strategies we have developed).

One weakness in our study is the inadequacy of the experimental design for studying the evolution of system-user interaction and user impressions over long periods of time. It is most reasonable to assume that the ECA may have a noticeable novelty effect on inexperienced users, which affects our observations. Nevertheless, observations from another line of research we are undertaking on ECA interfaces for children with motor disabilities suggest that (at least in certain contexts) the influence of the novelty effect should not be overstated. We will report results in future work.

We are now annotating the videos of the interactions in such a way as to make it easier to accumulate information on a variety of test parameters and even to share it with other research groups. Finally, using these videos, we plan to design tests to study the reactions of users to the emotional behaviour of the ECA, as a first step to modelling different types of users (e.g., extroverted/introverted, patient/irritable, etc).

We hope our work may help to show ways in which ECA technology can make a positive contribution to natural dialogue interfaces.

ACKNOWLEDGEMENTS

This work was carried out with the support of the European Union IST FP6 program, project COMPANION, IST-34434 and by the Spanish Ministry of Science and Technology under project TEC2006-13170-C02-01.

REFERENCES

[1] J. Cassell,, T. Bickmore, H. Vilhjálmsson, and H. Yan, More than just a pretty face: affordances of embodiment*, in Proceedings of the 5th international conference on Intelligent user interfaces*, pp. 52-59, ACM Press, 2000.

[2] S. J. Boyce, Spoken natural language dialogue systems: user interface issues for the future, *in Human Factors and Voice Interactive Systems,* D. Gardner-Bonneau Ed. Norwell, Massachusetts, Kluwer Academic Publishers: 37-62, 1999.

[3] D. McNeill, Hand and Mind: What Gestures Reveal about Thought. *The University of Chicago Press*, Chicago, 1992.

[4] P. Ekman, Facial Expression And Emotion, *American Psychologist*, 48(4), 384-392, 1993.
[5] M. Montepare, S.B. Goldstein, and A. Clausen, The iden-tification of emotions from gait information, *J. Nonverbal Behavior*, vol. 11, no. 1, pp. 33–42, Spring 1987.
[6] I. Poggi, C. Pelachaud and E.M. Caldognetto, Gestural Mind Markers in ECAs, *Gesture Workshop 2003*, pp 338-349, 2003.
[7] N. Leßmann, A. Kranstedt, and I. Wachsmuth, Towards a cognitively motivated processing of turn-taking signals for the embodied conversational agent Max, *Proceedings Workshop W12, AAMAS* 2004, New York, 57 - 65.
[8] R. Catrambone, Anthropomorphic agents as a user interface paradigm: Experimental findings and a framework for research, *in Proceedings of the 24th Annual Conference of the Cognitive Science Society*, pp. 166-171, Fairfax, VA, 2002.
[9] J. Xiao, Empirical Studies on Embodied Conversational Agents, *Ph.D. Dissertation*, Georgia Institute of Technology, Atlanta, GA, December 2006.
[10] COMPANIONS, European Commission Sixth Framework Programme Information Society Technologies Integrated Project IST-34434, http://www.companions-project.org/.
[11] M. McTear, Spoken Dialogue Technology: Towards the Conversational User Interface, *Springer*, 2004.
[12] T. Bickmore, J. Cassell, J. Van Kuppevelt, L. Dybkjaer, and N. Bernsen, (eds.), *Natural, Intelligent and Effective Interaction with Multimodal Dialogue Systems*, chapter Social Dialogue with Embodied Conversational Agents. Kluwer Academic, 2004.
[13] M. White, M. E. Foster, J. Oberlander, and A. Brown, Using facial feedback to enhance turn-taking in a multimodal dialogue system, *Proceedings of HCI International 2005*, Las Vegas, July 2005.
[14] S. Oviatt, and R. VanGent, Error resolution during multimodal humancomputer interaction, *Proc. International Conference on Spoken Language Processing*, 1 204-207, (1996).
[15] S. Oviatt, M. MacEachern, and G. Levow, Predicting hyperarticulate speech during human-computer error resolution, *Speech Communication*, vol.24, 2, 1-23, (1998).
[16] K. Hone, Animated Agents to reduce user frustration, in *The 19th British HCI Group Annual Conference*, Edinburgh, UK, 2005.
[17] S. Oviatt, Interface techniques for minimizing disfluent input to spoken language systems, *in Proc. CHI'94*, pp. 205-210, Boston, ACM Press, 1994.
[18] J. Cassell, Y.I. Nakano, T.W. Bickmore, C.L. Sidner, and C. Rich, Non-verbal cues for discourse structure, *in Proceedings of the 39th Annual Meeting on Association For Computational Linguistics*, 2001
[19] N. Chovil, Discourse-Oriented Facial Displays in Conversation, *Research on Language and Social Interaction*, 25, 163-194, 1992.
[20] A. Kendon, Conducting interaction: patterns of behaviour in focused encounters, *Cambridge Univer-sity Press*, 1990.
[21] J. Cassell and K.R. Thorisson, The power of a nod and a glance: envelope vs. emotional feedback in animated conversational agents. *Applied Artificial Intelligence*, vol.13, pp.519-538, (1999).
[22] R. San-Segundo, J.M. Montero, J. Ferreiros, R. Córdoba, and J.M. Pardo, Designing Confirmation Mechanisms and Error Recover Techniques in a Railway Information System for Spanish, *SIGDIAL*, Denmark, 2001.
[23] S. Oviatt, and B. Adams, Designing and evaluating conversational interfaces with animated characters, *Embodied conversational agents*, MIT Press: 319-345, 2000.
[24] H. Schaumburg, Computers as tools or as social actors: the users' perspective on anthropomorphic agents, *International Journal of Cooperative Information Systems*, pp 217-234, 2001.
[25] R. Catrambone, J. Stasko, and J. Xiao, ECA as user interface paradigm, From brows to trust: evaluating embodied conversational agents, *Kluwer Academic Publishers*, Norwell, MA, 2004.
[26] I. Bulyko, K. Kirchhoff, M. Ostendorf, and J. Goldberg, Error correction detection and response generation in a spoken dialogue system, *Speech Communication* 45, pp 271-288, 2005.
[27] S. Möller, P. Smeele, H. Boland, and J. Krebber, Evaluating spoken dialogue systems according to de-facto standards: A case study. Computer Speech & Language 21 (2007) 26-53.
[28] ITU-T P.851, Subjective Quality Evaluation of Telephone Services Based on Spoken. Dialogue Systems, *International Telecommunication Union (ITU)*, Geneva, 2003.
[29] K.Jokinen and T. Hurtig. User Expectations and Real Experience on a Multimodal Interactive System. In *INTERSPEECH-2006*, paper 1815-Tue2A3O.2.
[30] Nuance Communications' speech recognition technology, http://www.nuance.com.
[31] Haptek, http://www.haptek.com
[32] J. C. Lester, S. A. Converse, S. E. Kahler, S. T. Barlow, B. A. Stone and R. S. Bhogal, *The persona effect: affective impact of animated pedagogical agents*, in Proceedings of the SIGCHI conference on Human factors in computing systems, ACM Press New York, NY, USA, pp. 359-366, 1997 .
[33] A.B. Loyall, and J. Bates, Personality-rich believable agents that use language. In Johnson, W.L., Hayes-Roth, B., eds.: Proceedings of the First International Conference on Autonomous Agents (Agents'97), Marina del Rey, CA, USA, ACM Press (1997) 106–113.
[34] N.C. Krämer, G. Bente, and J. Piesk, The ghost in the machine. The influence of Embodied Conversational Agents on user expectations and user behaviour in a TV/VCR application. IMC Workshop (2003) 121-128.
[35] B. Reeves and C. Nass. The media equation: How people treat computers, television and new media like real people and places. CSLI Publications, Stanford,CA, 1996.
[36] B. López, Á. Hernández, D. Díaz, R. Fernández, L. Hernández, and D. Torre, Design and validation of ECA gestures to improve dialogue system robustness, Workshop on Embodied Language Processing, in the 45th Annual Meeting of the Association for Computational Linguistics, ACL, pp. 67-74, Prague, 2007.